Exhibit E

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 258, No. 17, Issue of September 10, pp. 10794-10804, 1983
Printed in U.S.A.

The Rat Casein Multigene Family

I. FINE STRUCTURE OF THE γ-CASEIN GENE*

(Received for publication, January 18, 1983)

Li-yuan Yu-Lee‡ and Jeffrey M. Rosen§

From the Department of Cell Biology, Baylor College of Medicine, Houston, Texas 77030

A region approximately 35 kilobase pairs (kb) in length containing the hormonally regulated rat γ-casein gene has been characterized by examining overlapping clones of genomic rat DNA obtained from two Charon 4A libraries. The entire γ-casein structural gene is contained in a single 17-kb phage clone. R-loop and restriction enzyme mapping analyses revealed that the γ-casein gene is approximately 15 kb long and is, therefore, 17.4 times larger than the mature γ-casein mRNA. The coding regions of the γ-casein gene are split into at least nine small segments, interspersed with long intervening sequences. Sequence analysis of the 5′ end of the γ-casein gene revealed the presence of a TATA sequence which may play a role in the initiation of gene transcription. The first exon is 44 nucleotides long and encodes part of the 5′ noncoding sequences of the γ-casein mRNA. The first intron was found to contain a short interspersed repeated DNA sequence which shares a 92% homology with a cloned rat repeated DNA sequence found at the 3′ end of several other rat genes. In addition, the γ-casein gene contains several families of highly repeated sequences interspersed throughout the intervening and flanking regions, including a family of evolutionarily conserved repeats. Thus, the γ-casein gene represents an unusually large and complex split mammalian gene.

The caseins are members of a small family of phosphoproteins which are secreted during lactation as large calcium-dependent aggregates termed micelles (Waugh, 1971) principally in response to the lactogenic hormones, prolactin and hydrocortisone (Topper, 1970). In the rat, the α-, β-, and γ-caseins constitute almost 80% of milk protein and have apparent molecular weights of 42,000, 25,000, and 22,000, respectively (Rosen *et al.*, 1975). The expression of the milk protein genes has been shown to be under complex multihormonal control both during normal mammary gland development (Rosen *et al.*, 1979) and in mammary gland explant cultures (Hobbs *et al.*, 1982). The kinetics of induction as well as the extent of mRNA accumulation in response to the hormones have also been found to be different for each of the three casein mRNAs (Hobbs *et al.*, 1982). This complex regulation may be attributed to the differential effects of hormones on both casein gene transcription and casein mRNA stability (Guyette *et al.*, 1979).

The examination of casein mRNA sequences by heteroduplex analysis (Richards *et al.*, 1981b) and direct sequencing (Blackburn *et al.*, 1982; Hobbs and Rosen, 1982) has indicated that there has been a considerable divergence among members of this small multigene family, except for three structural gene regions which are highly conserved: the signal peptide (Blackburn *et al.*, 1982; Rosen and Shields, 1980), the casein kinase recognition, and the 5′ noncoding sequences (Blackburn *et al.*, 1982; Hobbs and Rosen, 1982). It has been suggested by Hobbs and Rosen (1982) that the small family of functionally related caseins may have evolved by both intragenic and intergenic duplication of a primordial gene, which most likely contained the site of phosphorylation and calcium binding. Further support of the gene duplication hypothesis comes from genetic data which indicate that the bovine caseins occur as a gene cluster (Grosclaude *et al.*, 1973) and that all three casein genes are located on the same mouse chromosome (Gupta *et al.*, 1982).

In order to help elucidate the mechanisms by which hormones regulate the coordinate induction of these milk protein genes as well as to understand their mode of evolution, it was necessary to study initially their structure and genomic organization. This paper is the first in a series on the structure, organization, and evolution of the rat casein small multigene family. We report here the screening, restriction enzyme mapping, and partial sequencing of the large complex rat γ-casein gene. In the accompanying article (Johnson *et al.*, 1983), the state of DNA methylation of certain sites within the γ-casein as well as the β-casein genes is examined as a function of the expression of these genes.

EXPERIMENTAL PROCEDURES

Materials—All restriction enzymes were purchased from Bethesda Research Laboratories, New England Biolabs, or Boehringer Mannheim and used according to the specifications of the suppliers. DNA polymerase I was from Boehringer Mannheim. DNA polymerase I Klenow fragment was from New England Biolabs. Reverse transcriptase was the kind gift of Dr. J. Beard, Life Sciences Inc., St. Petersburg, FL. T4 DNA ligase, T4 polynucleotide kinase, and bacterial alkaline phosphatase were from Bethesda Research Laboratories.

Library Screening and Phage DNA Preparation—Two rat genomic libraries, prepared from random partial *Eco*RI (Sargent *et al.*, 1979) or *Hae*III digestions of Sprague-Dawley rat DNA and cloned into Charon 4A phage (Blattner *et al.*, 1977), were kindly provided by Drs. T. Sargent, B. Wallace, and J. Bonner, and Drs. L. Jagodzinski and J. Bonner, respectively (California Institute of Technology). The libraries plated on *Escherichia coli* LE392 were screened according to the procedure of Maniatis *et al.* (1978), using either a γ-casein cDNA probe (*Pst*I-released 903-bp[1] insert) (Richards *et al.*, 1981a) or subsequently a 4.1-kb exon-containing *Bam*HI subgenomic fragment. The top agar layer routinely contained 0.7% agarose. ^{32}P-labeled probes were nick-translated according to the procedures of Rigby *et*

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Recipient of National Institutes of Health Fellowship CA 06645.

§ Supported by Grant CA 16303 from the National Institutes of Health. To whom reprints requests should be addressed.

[1] The abbreviations used are: bp, base pair; kb, kilobase pair, 1 × SSC = 0.15 M NaCl, 0.015 M Na citrate; 1 × Denhardt = 0.02% Ficoll, 0.02% polyvinylpyrrolidone, 0.02% bovine serum albumin.

al. (1977) to 1–5 × 10^8 cpm/μg of specific activity. Labeled probes were routinely denatured in the presence of 0.5–1 mg of sheared *E. coli* DNA and 50–100 μg of pBR322 DNA and prehybridized for 5 min at 68 °C to reduce background hybridization signals. It was critical to remove even small amounts of contaminating plasmid sequences from the probes prior to screening since the *Hae*III library is contaminated with pBR322 sequences.[2] Filters containing recombinant plaques were screened according to the procedure of Benton and Davis (1977), as modified by Woo *et al.* (1978), baked at 68 °C for a minimum of 4 h, and prehybridized overnight in 6 × SSC and 2 × Denhardt (Denhardt, 1966). Filters were hybridized with 500,000 cpm/filter of ^{32}P-labeled DNA and washed in 2 × SSC, 0.1% sodium dodecyl sulfate. After the second round of plaque purification, single plaques were transferred with a toothpick to a plate containing a fresh lawn of bacteria, permitting the screening of a large number of plaques on only a few filters.

Large scale phage DNA preparations were carried out as described by Maniatis *et al.* (1982) except that the infected cells were shaken vigorously in 2-liter Bellco baffled flasks to ensure full aeration. Phage minilysis preparations were isolated as described by Leder *et al.* (1977) with the following modifications. After lysis, the supernatant was treated with 10–50 μg/ml of DNase I for 30 min at room temperature, followed by a 30-min incubation at 37 °C with 50–100 μg/ml of proteinase K, 0.2% sodium dodecyl sulfate, 50 mM Tris-HCl (pH 7.4), 20 mM Na_2EDTA. The lysate was extracted twice with phenol and once with chloroform, precipitated with ethanol, resuspended in water, and digested with 20 μg/ml of RNase A (pretreated for 15 min at 90 °C). One half to one fifth of the sample was digested with the appropriate restriction enzyme (1–5 units for 2–4 h) and analyzed on a horizontal 1% agarose gel in Tris acetate buffer (Maniatis *et al.*, 1978).

Gel Transfer and Hybridization—DNA fragments were transferred to nitrocellulose filters either by the method of Southern (1975) or by the bidirectional transfer method of Smith and Summers (1980). Transfers employing the latter method were completed between 1 and 3 h and baked for not more than 8 h. Filters containing cloned DNA fragments were hybridized routinely with 6–12 × 10^6 cpm of ^{32}P-labeled probe for 12 h. Filters containing total rat genomic DNA were hybridized with 40–50 ng of ^{32}P-labeled probe/lane containing 15 μg of DNA in the presence of 10% dextran sulfate according to the procedure of Wahl *et al.* (1979). Filters containing cloned DNA were routinely re-used up to four times by washing in 10 mM Tris base at 68 °C for 1 h according to Griffin-Shea *et al.* (1980), followed by a rinse in water and prehybridization as described.

Plasmid Subcloning—DNA fragments from phage clones were subcloned into either pBR322, pBR325 (Bolivar, 1978; Prentki *et al.*, 1981), or pUC8 (Vieira and Messing, 1982) plasmids. After the appropriate enzyme digestion of genomic DNA in phage clones, DNA fragments were isolated from an agarose gel by electroeluting either directly into a dialysis bag (>10 μg/band) or onto DEAE-81 paper (<10 μg/band) according to Dretzen *et al.* (1981). DEAE-bound DNA (from 100 bp up to 5 kb) was washed twice with 50 mM NaCl and eluted with two 30-min incubations of 1 M NaCl or 1 M NaCl followed by 1 M Na acetate (pH 4.5). The eluate was clarified by centrifugation in a microfuge, precipitated directly with 2.5 volumes of ethanol, frozen at −80 °C for 15 min, pelleted, and resuspended in water. DNA isolated using this procedure was found to be suitable for subsequent ligation, subcloning, nick translation, end labeling, and sequencing.

Linearized plasmid vectors were treated with bacterial alkaline phosphatase (20 units of bacterial alkaline phosphatase/pmol of DNA end) as described by Chaconas and van de Sande (1980). Ligation reactions were performed as described by Dugaiczyk *et al.* (1975) except that the molar end ratio of plasmid vector to insert DNA was kept at 1:1 to minimize multiple insertions into a single vector. The ligation mixtures were routinely heated at 68 °C for 5 min to dissociate all sticky ends and quick chilled on ice before T4 DNA ligase was added to 50 units/pmol of plasmid DNA end. All ligation reactions were incubated in a refrigerator at 8–12 °C overnight. *E. coli* RRI cells were used for transformation with pBR322 and pBR325 plasmid vectors, while *E. coli* JM83 cells (Vieira and Messing, 1982) were used with pUC8 vectors. Cells were made competent for transformation by an overnight incubation at 4 °C in 0.1 M $CaCl_2$ as described by Dagert and Ehrlich (1979) and Norgard *et al.* (1978). Transformation was carried out as described by Dagert and Ehrlich (1979), except that for pUC8 cloning, ampicillin-supplemented (50 μg/ml) L plates (100 × 15 mm) were first spread evenly with 40 μl of 2% 5-bromo-4-chloro-3-indolyl-β-D-galactoside in dimethylformamide before plating. Recombinant pUC8 plasmids were routinely retransformed into RRI cells in order to maximize the yield of plasmid DNA. Plasmid DNAs were analyzed by the rapid plasmid minilysate method of Holmes and Quigley (1981). Lysates prepared from 3-ml overnight cultures were routinely boiled for 2.5 min to maximize plasmid yield which was usually sufficient for three to five restriction digestions. Lysates prepared by this method were free of chromosomal DNA as well as RNAs which may interfere with enzyme digestions. For rapid analysis, restriction enzyme-digested plasmid subclones were electrophoresed for 30 min at 100–150 V (about 150–200 mA) on a 1% agarose minigel formed on a glass slide. Using this method, as little as 10 ng of DNA (0.5–1-kb range) can be easily visualized with ethidium bromide staining.

Large scale plasmid DNA preparation was performed as described by Norgard *et al.* (1979) employing 1 mg/ml of uridine and either chloramphenicol (250 μg/ml) for pBR322 and pUC8 or spectinomycin (300 μg/ml) for pBR325 amplification. After centrifugation, cells were washed once with 25 mM Tris-HCl (pH 7.4), 10 mM Na_2EDTA and, without any freeze-thawing, lysed according to Katz *et al.* (1973). Following CsCl banding, plasmid DNA was treated with DNase-free RNase A, followed by proteinase K as described previously (Blackburn *et al.*, 1982; Richards *et al.*, 1981b).

R-loop Analysis—Phage clones (10 μg/ml) containing γ-casein DNA were hybridized to a γ-casein mRNA-enriched Sepharose 4B column fraction (30 μg/ml) (Rosen, 1976) in 70% formamide, 0.3 M NaCl, 10 mM Tris-HCl (pH 7.4), 10 mM Na_2EDTA. The mixture was denatured for 2 min at 80 °C and hybridized at 52 °C for 3h, followed by an incubation at 42 °C for 1–3 h, cooling to room temperature, and spreading in 70% formamide, 100 mM Tris-HCl (pH 8.5), 10 mM Na_2EDTA onto copper grids (300 mesh, EBTEC, Agawam, MA). The grids were stained in uranyl acetate, shadowed with platinum-paladium, and examined under a JEOL 100CX electron microscope.

End Labeling and Sequencing of DNA—DNA fragments were labeled either at their 3′ ends with [α-^{32}P]dNTP using DNA polymerase I Klenow fragment according to Maniatis *et al.* (1982) or at their 5′ ends with [γ-^{32}P]ATP after treatment with bacterial alkaline phosphatase followed by T4 polynucleotide kinase according to Maxam and Gilbert (1980). The end-labeled fragments were then digested with a second enzyme to generate singly end-labeled fragments, resolved by electrophoresis on an acrylamide gel (from 4 up to 8%) in Tris borate buffer (Maxam and Gilbert, 1980), and isolated according to Maxam and Gilbert (1977). Isolated DNA fragments were partially digested with restriction enzymes according to Smith and Birnsteil (1976) to localize restriction sites. Sequencing of end-labeled fragments was performed by using the method of Maxam and Gilbert (1980) on 8 and 20% acrylamide/urea gels. Computer analysis of sequence data was performed as described previously (Blackburn *et al.*, 1982).

5′ Extension of mRNA—The cloning (Richards *et al.*, 1981, a and b) and sequencing (Hobbs and Rosen, 1982) of the rat γ-casein cDNA (pCγ41) have been reported previously. This γ-casein cDNA clone has been shown to be missing 35 nucleotides of the 5′ mRNA sequence (Hobbs and Rosen, 1982). In order to localize the 5′ end of the γ-casein gene in the various phage clones, it was necessary to prepare a DNA fragment corresponding to the 5′-most mRNA sequence to be used as probe. A 5′ primer DNA was prepared from the γ-casein cDNA by first digesting with *Pst*I to release the insert and then redigesting with *Acc*I and *Sau*3A to generate a 104-nucleotide *Sau*3A/*Acc*I primer fragment (see Fig. 4 of Hobbs and Rosen, 1982) which was isolated from an 8% acrylamide gel. Primer extension was performed by a modification of the method of Smith (1980) as follows.

Primer DNA was hybridized with 3 μg of total poly(A) RNA isolated from an 8-day lactating rat mammary gland (Rosen, 1976), boiled for 2 min at 90 °C, and annealed for 20 min at 68 °C. The primer/RNA template mixture was then added to a final volume of 25 μl of cDNA synthesis buffer with a final concentration of 5 mM $MgCl_2$, 10 mM dithiothreitol, 500 μM concentration each of all four deoxynucleotide triphosphates, 4 mM Na pyrophosphate, 50 μCi each of [α-^{32}P]dATP and [α-^{32}P]dCTP, and 1.5 units/μl of reverse transcriptase. The cDNA synthesis reaction was incubated for 20 min at 46 °C, stopped by heating for 2 min at 68 °C, and chromatographed on Bio-Gel P-60 to remove unincorporated triphosphates. The products were precipitated by ethanol and resuspended in water, and the RNA component was hydrolyzed by boiling in 0.3 N NaOH for 15 min. Labeled single-stranded 5′ extended cDNA was neutralized and added directly to filters for hybridization for 48 h at 68 °C.

Recombinant DNA Safety—All experiments involving recombinant

[2] L. Jagodzinski, and J. Bonner, personal communication.

DNA were performed in accordance with the "National Institutes of Health Guidelines for Research Involving Recombinant DNA Molecules."

RESULTS

Mapping the Rat γ-Casein Gene Region—Previous restriction enzyme mapping studies and gene dosage experiments (Johnson *et al.*, 1983) using total rat DNA indicated that the γ-casein gene is quite large and that there is only a single copy of the gene in the rat genome. It was, therefore, estimated that several overlapping clones would be needed to encompass the entire γ-casein gene and its flanking sequences. Two independent rat DNA libraries were used to screen for the γ-casein gene since it seemed possible that an area with multiple closely spaced *Eco*RI sites might not be represented in the *Eco*RI library. Primary screening of 9,000–10,000 plaques/plate with a 99% full length cloned γ-casein cDNA probe (Richards *et al.*, 1981b) gave weak signals as shown in Fig. 1*A* (*lane 1*). To increase the signal to noise ratio, we employed a larger subgenomic fragment, a 4.1-kb *Bam*HI fragment (see below) as a probe. As shown in Fig. 1*A* (*lane 3*), the signal in the primary screen was greatly enhanced and was comparable in intensity to that in a secondary screen using the cDNA probe (Fig. 1*A*, *lane 2*).

A total of seven and nine γ-casein specific clones were obtained from the *Eco*RI and *Hae*III libraries, respectively. The clones from the *Eco*RI library displayed only two different *Eco*RI digestion profiles; therefore, only two clones, λγ1 and λγ3, were analyzed further (Fig. 1*B*, *lanes 9* and *10*). The λγ1 insert DNA contains a single *Eco*RI site, splitting the DNA into 9.7- and 7.3-kb fragments, both of which contain γ-casein-coding sequences (Fig. 1*C*, *lane 9*). To establish the transcriptional orientation of the DNA in the λγ1 clone, 5′-*versus* 3′-specific cDNA probes were generated by digesting the γ-casein cDNA at a unique *Sst*I site (Richards *et al.*, 1981, a and b). It was observed that the 5′-specific cDNA probe hybridized to both *Eco*RI fragments in λγ1 but that the 3′ probe hybridized only with the 7.3-kb *Eco*RI fragment, thereby establishing the orientation in the γλ1 clone (data not shown). The λγ3 clone overlaps the λγ1 clone through the 7.3-kb fragment and extends in the 3′ direction (Fig. 1, *B* and *C*, *lanes 10*). Similarly, several clones from the *Hae*III library were also found to be identical with each other (Fig. 1, *B* and *C*). For example, the *Eco*RI-digested phage DNA profile shown in *lane 4* is identical with those in *lanes 6* and *7*. Only four clones were further analyzed, λγ9, λγ10, λγ7, and λγ8.

The results of these and numerous other restriction enzyme-mapping experiments of these phage clones (data not shown) were used to generate a map of the rat γ-casein gene region as shown in Fig. 2. The overlapping phage clones are aligned according to the *Eco*RI, *Bam*HI, and *Sst*I restriction sites. The entire γ-casein gene region spans about 35 kb of the rat genome. Approximately 9 kb of 5′ flanking sequences are found in the λγ9 and λγ10 clones, while about 11 kb of the 3′ flanking DNA are found in the λγ3 clone. Several restriction site polymorphisms were observed in λγ9, λγ10, and λγ7 DNAs, most notably in the flanking regions but also within the 5′ intervening sequences. We suggest that the anomalies are due to polymorphism in the Sprague-Dawley rat DNA used for library construction, as has been reported previously (Esumi *et al.*, 1982). However, the insertion in both λγ9 and λγ10 as well as the deletion in λγ7 may have resulted from recombination during cloning and/or during propagation of phage DNA.

In order to demonstrate that the λγ1 clone contains a bona fide γ-casein gene sequence, it was necessary to compare the

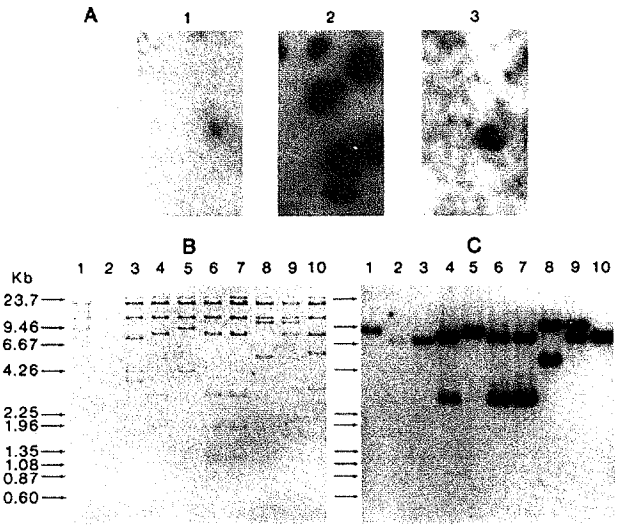

FIG. 1. **Phage clone screening and identification by minilysis and DNA blot analysis.** Recombinant phage clones were screened from both a partial *Eco*RI and a partial *Hae*III rat DNA library using either a cloned γ-casein cDNA insert (*A*, *lanes 1* and *2*) or a 4.1-kb subgenomic *Bam*HI fragment (*A*, *lane 3*) as probe (see Fig. 5 for position). Positive clones from the *Hae*III library which were screened with the 4.1-kb fragment were identified by a small scale minilysis analysis. The lysates were digested with *Eco*RI (10 units/0.3 ml of lysate), electrophoresed on a 1% agarose gel (*B*, ethidium bromide profile), transferred to nitrocellulose, and hybridized with a ^{32}P-labeled γ-casein cDNA probe (*C*, autoradiogram). The recombinant phage clones were designated as follows for *B* and *C*: λγ9 (*lane 3*), λγ10 (*lane 5*), λγ7 (*lane 7*), λγ8 (*lane 8*), λγ1 (*lane 9*), λγ3 (*lane 10*). *Lane 2* was underloaded. The *first three bands* in each lane correspond to the λ phage arms of the Charon 4A vector: joint arms due to the *Eco*RI sticky ends, left arm and right arm, respectively. *Arrows* indicate the positions of molecular weight standards.

restriction enzyme map of the γ-casein gene in total genomic rat DNA with that in λγ1 cloned DNA (Fig. 3). As shown before, λγ1 contains two *Eco*RI fragments (Fig. 3, *A* and *B*, *lane 1*). The γ-casein gene sequences found in total rat liver DNA are also contained in two *Eco*RI fragments of slightly larger sizes, about 13 and 9.3 kb in length (Fig. 3*C*, *lane 1*). The 14.3- and faint 8.4-kb *Bam*HI fragments in genomic DNA (Fig. 3*C*, *lane 2*) corresponded, respectively, to the 3′ 27- and 5′ 7.5-kb *Bam*HI fragments still attached to phage arms in the λγ1 clone (Fig. 3*B*, *lane 2*). The internal 4.1- and 2.3-kb *Bam*HI fragments were as predicted from the λγ1 DNA. These results demonstrated that there were no major rearrangements of the γ-casein gene sequences during either cloning and/or propagation of the λγ1 DNA.

Fine Structure of the γ-Casein Gene—The above results suggested that the λγ1 clone contains the majority of the γ-casein-coding sequences and may contain the entire γ-casein gene. We therefore characterized it further by R-loop analysis, subcloning, and detailed restriction enzyme mapping. An R-loop analysis was carried out between the λγ1 DNA and γ-casein mRNA to delineate exon and intron positions as shown in Fig. 4. Electron microscopic examination of the hybrid molecule revealed at least nine small coding regions separated by a minimum of nine large loops which corresponded to intervening and/or flanking sequences. The results of numerous mapping studies (data not shown) and the R-loop data were used to construct a more detailed map of the rat γ-casein gene (Fig. 5). The size of the rat γ-casein gene is approximately 15 kb. Its coding sequences are separated by multiple intervening sequences, the largest of which is about 3 kb. Since mature γ-casein mRNA is 869 nucleotides in length

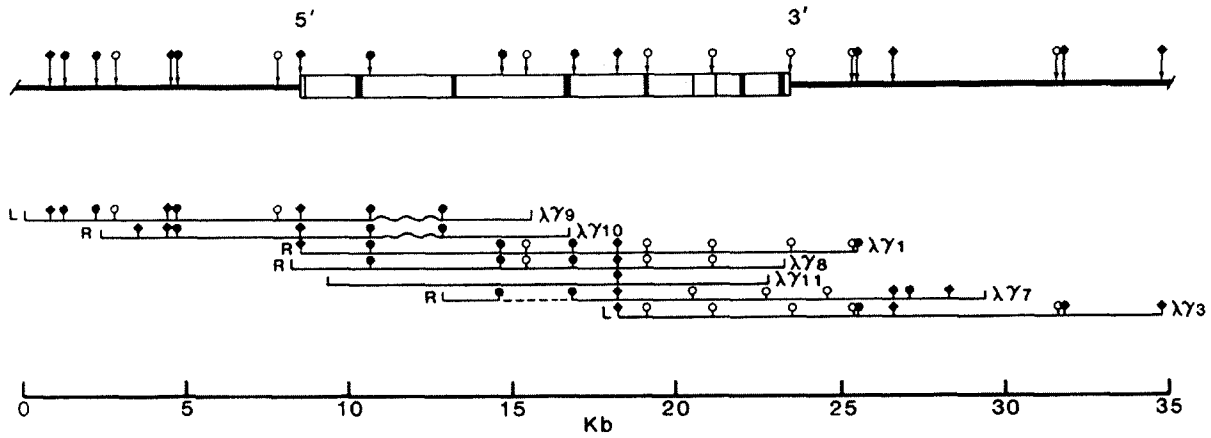

FIG. 2. **Map of the rat γ-casein gene region.** Overlapping γ-casein-specific phage clones are shown in the 5′ to 3′ orientation. *L* and *R* indicate the orientation of the left and right arms, respectively, of the Charon 4A vector. *Solid boxes* represent exons, *open boxes* represent intervening sequences, and the *solid line* represents flanking sequences. The *dashed line* denotes a 2.2-kb deletion in clone λγ7. The *wavy line* denotes a 2.3-kb insertion in clones λγ9 and λγ10. Clone λγ10 has been mapped with only *Eco*RI and *Bam*HI. Clone λγ11 has only been characterized with *Eco*RI and *Msp*I (data not shown). Clones λγ1 and λγ3 have true *Eco*RI ends; all of the other clones have artificial *Eco*RI linkers as ends (Maniatis *et al.*, 1982). ◆, *Eco*RI; ●, *Bam*HI; ○, *Sst*I. The *scale* is shown in kilobase pairs.

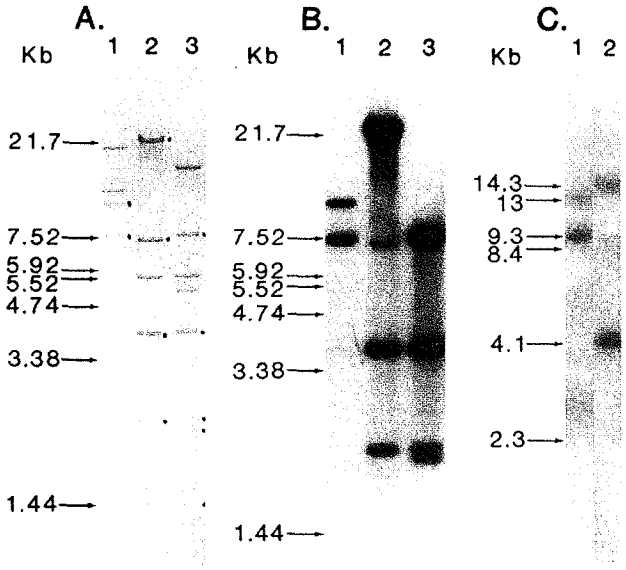

FIG. 3. **Hybridization of γ-casein cDNA to genomic and cloned rat DNA.** λγ1 clone DNA was digested with *Eco*RI (*lane 1*), *Bam*HI (*lane 2*), and both *Eco*RI and *Bam*HI (*lane 3*), electrophoresed on a 1% agarose gel (*A*), and hybridized with a ^{32}P-labeled γ-casein cDNA probe to localize all of the exon-containing fragments (*B*). *C*, total rat liver DNA (15 μg/lane) was digested with *Eco*RI (*lane 1*) and *Bam*HI (*lane 2*) and analyzed on an 0.8% agarose gel. *Dotted bands* in *A* denote γ-casein DNA-containing fragments. Molecular weight sizes are designated by the *arrows*.

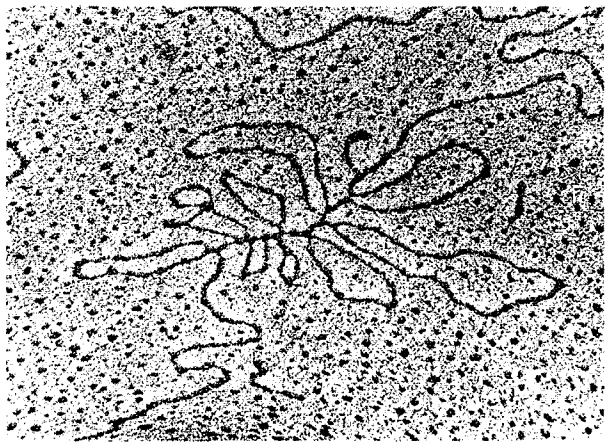

FIG. 4. **R-loop analysis of λγ1 clone.** λγ1 DNA was denatured and hybridized to Sepharose 4B fraction enriched γ-casein mRNA as described under "Materials and Methods." The R-loops were used to localize approximately the positions of exons and intervening sequences in the γ-casein gene as shown in Figs. 2 and 5.

(Hobbs and Rosen, 1982), the 15-kb γ-casein gene is, therefore, 17.4-fold larger than the mRNA it encodes. The second *Sst*I site in the gene splits an exon which agrees well with the single *Sst*I site at position 534 located in the middle of the γ-casein cDNA sequence (Richards *et al.*, 1981b). Several potential methylation sites, as determined by *Msp*I and *Hha*I digestions, were found in the intervening and flanking sequences (see Johnson *et al.*, 1983).

For fine mapping and sequencing, the γ-casein gene was subcloned in fragments spanning its entire length into pBR322, pBR325, or pUC8 vectors, using either single *Eco*RI and *Bam*HI or double *Eco*RI and *Bam*HI cloning sites. Two subclones in particular were analyzed further and used to define the 5′ and 3′ ends of the γ-casein gene: pλγ2.1, which contains the 2.1-kb *Eco*RI and *Bam*HI fragment at the 5′ end of the gene, and pλγ3.4, which contains the 3′ region of the gene.

Defining the 5′ End of the γ-Casein Gene—The 5′ end of the γ-casein gene was localized in a 0.63-kb *Eco*RI/*Hind*III fragment in the subclone pλγ2.1 as shown in Fig. 6. A 5′ primer-extended single-stranded cDNA (about 44 bp in length) (see "Materials and Methods") was used as a hybridization probe. As seen in Fig. 6*B*, the 5′-most exon is contained in the 3.28-kb *Hind*III fragment (*lane 1*) which was reduced to a 0.63-kb fragment when a double digestion was performed with *Eco*RI (*lane 2*). Similarly, the 5′-most exon is contained in a 0.87-kb *Eco*RI/*Rsa*I fragment (*lane 4*). A second exon was also localized to a 0.29-kb *Rsa*I fragment (*lanes 3* and *4*) in this subclone.

Defining the 3′ End of the γ-Casein Gene—The next step

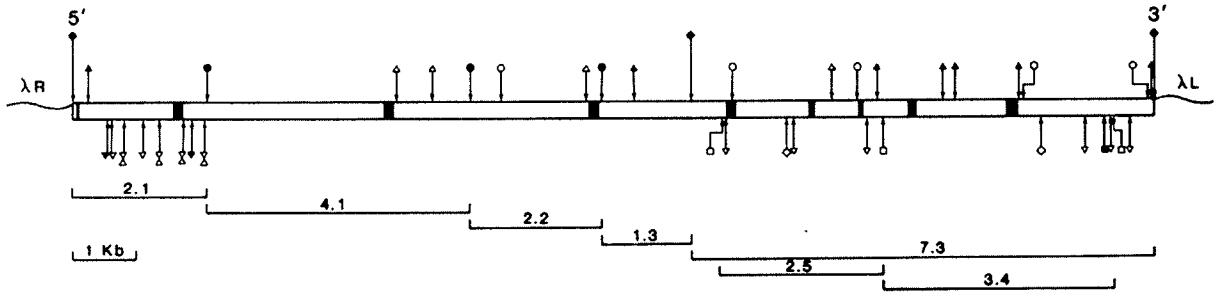

FIG. 5. **Map of the 15-kb rat γ-casein gene.** The entire rat γ-casein gene shown in the 5′ to 3′ orientation is localized in the 17-kb λγ1 phage clone. *Solid boxes* show the sizes and positions of the γ-casein mRNA sequences, and *open boxes* represent either intervening or flanking sequences in the clone. Fragments spanning the entire gene have been subcloned into plasmid vectors. The size of the subcloned subgenomic fragments (in kilobase pairs) are shown *below* the map. The subclones were designated as pλγ2.1, pλγ4.1, and so forth. *Downward pointing arrows* indicate restriction enzyme positions mapped with phage λγ1 DNA. *Upward pointing arrows* indicate restriction sites localized by mapping with subcloned DNAs. Note that the 5′-most *Msp*I site is composed of two *Msp*I recognition sequences which are separated by only three nucleotides, as determined by DNA sequencing (see Fig. 8*B*, second *g* region). λ*R* and λ*L* refer to the right and left arms of the Charon 4A vector. ▼, *Acc*I; △, *Hha*I; □, *Pst*I; ●, *Bam*HI; ▽, *Hind*III; ⧖, *Rsa*I; ◇, *Bgl*II; ■, *Kpn*I; ○, *Sst*I; ◆, *Eco*RI; ▲, *Msp*I.

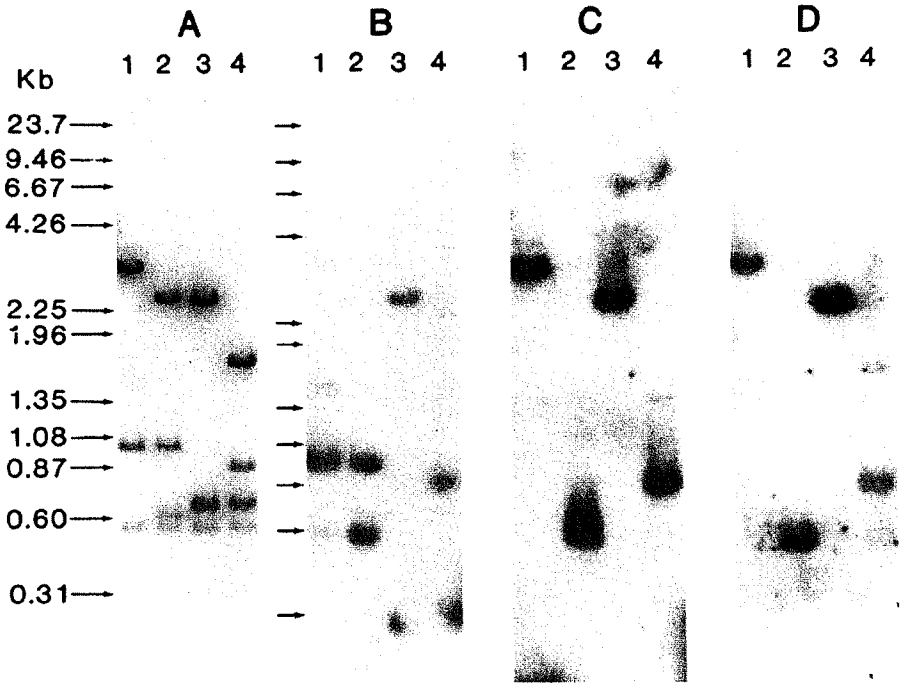

FIG. 6. **Defining the 5′ end of the γ-casein gene.** *A*, ethidium bromide pattern of subclone pλγ2.1 DNA digested with *Hind*III (*lane 1*), *Hind*III plus *Eco*RI (*lane 2*), *Rsa*I (*lane 3*), and *Rsa*I plus *Eco*RI (*lane 4*) and electrophoresed on a 1.5% agarose gel. *B*, after bidirectional transfers, the bottom filter was hybridized to a 5′ primer-extended single-stranded γ-casein cDNA probe to localize the 5′-most exon. To localize the positions of repeated sequences (see also Fig. 9), hybridizations were carried out with a ^{32}P-labeled total rat liver DNA probe at either 52 °C (*C*) (reused bottom filter) or 68 °C (*D*) (top filter). *Arrows* indicate the positions of molecular weight standards.

was to localize the 3′ end of the γ-casein gene (Fig. 7). Subclone pλγ3.4 was digested with *Msp*I and *Pst*I (Fig. 7*A*, *lane 1*) to generate insert bands 2.05, 1.1, and 0.63 kb long. According to the λγ1 map in Fig. 5, the last exon should be contained in a 1.1-kb *Msp*I fragment or a 1.7-kb *Sst*I/*Pst*I fragment (Fig. 7*A*, *lane 2*). Hybridization results using a specific 3′ end cDNA probe indicated that, indeed, the last 59 nucleotides of the 3′ noncoding sequence of the γ-casein mRNA was located in the 1.1-kb *Msp*I fragment (Fig. 7*A*, *lane 3*). Additional support for the location of the 3′-most exon came from a hybridization experiment using the 3′ end phage clone λγ3 (Fig. 7*B*). Only one band in each lane hybridized with the 3′ probe, demonstrating satisfactorily that the rest of the 11 kb of DNA sequences in λγ3 contained only 3′ flanking DNA sequences. Thus, the 15-kb γ-casein gene is located in its entirety in the 17-kb λγ1 clone.

Sequence Analysis of the 5′ End of the γ-Casein Gene—From the results in Fig. 6, it seemed likely that the 5′-most exon is located close to the *Eco*RI site at the end of pλγ2.1, since double digestions performed with *Eco*RI always generated a subfragment which hybridized with the 5′-primer-extended probe. Strategies were therefore devised to sequence the 5′ region of the γ-casein gene in pλγ2.1, as shown in Fig. 8*A*. The nucleotide sequence of the 5′-most 0.6-kb *Eco*RI/*Acc*I fragment is presented in Fig. 8*B*. The 5′ end of the γ-casein mRNA is encoded by an exon located about 100 nucleotides from the 5′ *Eco*RI site, which contains the first 44 nucleotides of the 5′ noncoding region of the mRNA. The putative mRNA CAP site was located on residue A at nucleotide +1 which agrees well with the γ-casein mRNA sequence data obtained by Hobbs and Rosen (1982). Further upstream in the 5′ flanking region is an AT-rich sequence from nucleotide −29 to −23, TTTAAAT, bounded on both sides by GC-rich residues, and is identified as a possible Goldberg-Hogness (TATA) sequence (Breathnach and Chambon, 1981). A second sequence from nucleotide −86 to −80, CGAATGT, dis-

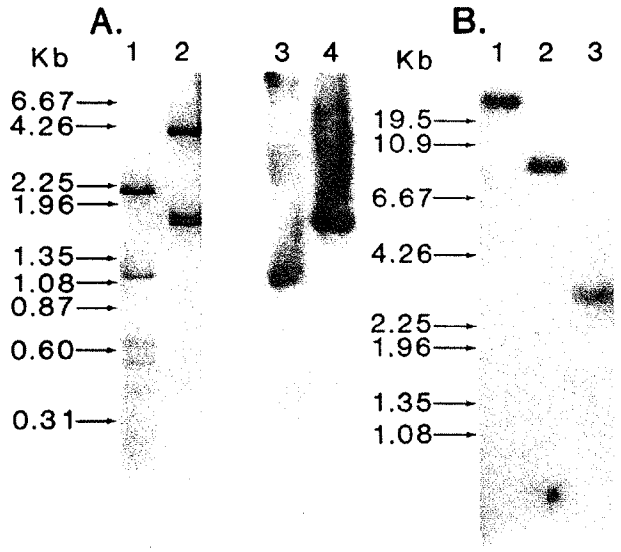

FIG. 7. **Defining the 3′ end of the γ-casein gene.** A subclone of the 3′ end of the γ-casein gene (pλγ3.4) and the 3′-most phage clone (λγ3) were used to delineate the 3′ end of the gene. *A*, plasmid pλγ3.4 DNA (see Fig. 5) was digested with *Msp*I plus *Pst*I (*lane 1*) and *Sst*I plus *Pst*I (*lane 2*) and electrophoresed on a 2% agarose gel. *Lanes 3* and *4*, autoradiograms. *B*, phage λγ3 DNA was digested with *Bam*HI (*lane 1*), *Eco*RI plus *Bam*HI (*lane 2*), and *Sst*I (*lane 3*). To prepare a specific 3′ end cDNA probe, the *Pst*I-released γ-casein cDNA insert was digested with *Rsa*I and end-labeled at the 3′ terminus, and the 116-nucleotide *Rsa*I/*Pst*I fragment containing 59 nucleotides of the 3′ noncoding γ-casein mRNA sequence and 38 As of the poly(A) tail plus 19 Cs which were part of the linkers for the *Pst*I cloning site (see Fig. 4 of Hobbs and Rosen, 1982) was isolated from a 5% acrylamide gel and used as probe. *Arrows* indicate positions of molecular weight standards.

plays a weak homology with the canonical CAAT sequence which has been observed in some other eucaryotic genes (Benoist *et al.*, 1980). The sequence of the first intron begins with the dinucleotide GT, which conforms to the consensus sequence AG/GT for the 5′ exon/intron splice junctions of a large number of eucaryotic genes (Breathnach and Chambon, 1981; Mount, 1982).

Repeated Sequences in the γ-Casein Gene: Repeats in the First Intron—The first intervening sequence contains a number of interesting features (Fig. 8*B*). A 282-bp sequence starting after the gap at nucleotide 151 is observed to share a 92% homology with a repeated DNA sequence (psubK) (Sargent, 1981), reiterated 100,000 times/haploid rat genome, which has been subcloned from the 3′ flanking region of the rat serum albumin gene (Sargent, 1981). This high degree of homology reflects only 23 single-bp substitutions and 2 single-bp deletions found in the psubK sequence as compared with the γ-casein intron I sequence. Furthermore, the homology observed with the psubK sequence is in the opposite orientation to the direction of transcription of the γ-casein gene. The first one third of the homologous intron I repeat corresponds to only 5% of the actual psubK repeat sequence, while the last two thirds are homologous with a unique DNA sequence in the 3′ flanking region of the rat serum albumin gene.

Several short direct repeats are found in the intron region immediately 3′ to exon I (Fig. 8*B*). For example, at nucleotide 54 is the sequence CATATT which is repeated again around nucleotide 87. The sequence GTTAAC at nucleotide 102 is repeated again at nucleotide 138. Furthermore, this sequence is part of a longer sequence of 14 nucleotides, CAAAGTATGTTAAC, beginning at nucleotide 94, which shares an 81% homology with a 16-bp sequence, CCAAGCTATAGTTAAC, beginning at nucleotide 128. Interestingly, between this directly repeated sequence is the sequence AATAAA (Proudfoot and Brownlee, 1976), which may form part of the recognition site for polyadenylation as demonstrated for SV40 late mRNAs (Fitzgerald and Shenk, 1981). In addition, this AATAAA sequence is also found at the more distal boundary of the intron I repeat sequence. Thus, the first intron of the γ-casein gene contains a repetitive element, part of which is reiterated 100,000 times elsewhere in the rat genome.

Other Repeated DNA Sequences in the γ-Casein Gene Region—In addition to the repetitive element sequenced in the first intron, other repeated DNA sequences were also observed throughout the γ-casein gene region. The existence of repeated DNA sequences in the γ-casein gene was initially suggested by a study in which λγ1 DNA was nick-translated and hybridized with total rat liver DNA. A smear of radioactivity was obtained instead of discrete bands, indicating that repeated sequences are present in the λγ1 DNA probe and that these sequences are not located in tandem but are dispersed throughout the rat genome (data not shown). To further localize the positions of these repeated sequences, total rat liver DNA was nick-translated and hybridized with cloned λγ1 DNA. A representative study is shown in Fig. 9. Under conditions of reduced stringency (2 × SSC, 52 °C), three bands displayed strong hybridization signals (Fig. 9*A*, *lane 4*), indicating the presence of repeated sequences. When the stringency was increased to 68 °C, the intensity of the hybridization signals of the two upper bands was reduced (Fig. 9*A*, *lane 6*). The strong hybridization signals observed in the 5′ end subclone, pλγ2.1, at either 52 or 68 °C in Fig. 6, *C* and *D*, were likewise due to the presence of a highly repeated DNA sequence as confirmed by direct DNA sequencing (Fig. 8*B*).

The repeated sequences are dispersed within and surrounding the γ-casein gene, as illustrated by the experiment shown in Fig. 9*B*, where λγ1 DNA was digested with various restriction enzyme combinations and hybridized at 52 °C. Similar experiments were carried out using the cloned psubK repeat DNA (Sargent, 1981) as a probe under conditions of either low or high stringency (data not shown). The results of numerous experiments localizing repeated sequences in the γ-casein gene are summarized in the diagram in Fig. 9*C*. Repeated sequences are found throughout the intervening and flanking sequences in the γ-casein gene, except for a 2-kb intron region towards the middle of the gene. There seem to be several families of repeated sequences as demonstrated by the different intensities of hybridization using total DNA as well as the cloned rat repeat DNA as probes. Moreover, three regions in the γ-casein gene were found to contain evolutionarily conserved repeated DNA sequences which were first reported by Miesfeld *et al.* (1981) and were observed only when carrier salmon sperm DNA was not used in the hybridizations (data not shown) (Miesfeld *et al.*, 1981). The 2-kb deletion observed in the λγ7 DNA in Fig. 2 may be attributed to the highly repeated DNA (within the second *e* region in Fig. 9*C*) whose presence may have contributed to the instability of this region.

In addition to the presence of highly repeated sequences in the γ-casein gene, some introns appear to contain sequences that are reiterated only within and/or around the gene itself. For example, pλγ2.1, the subclone containing the 5′ end of the gene, contains a sequence that is also repeated in a 5 kb-*Eco*RI fragment in the 3′ flanking region (data not shown). Similarly, pλγ4.1, the subclone in the 5′ portion of the gene which was also used for screening the *Hae*III rat library, contains a sequence that is repeated immediately downstream

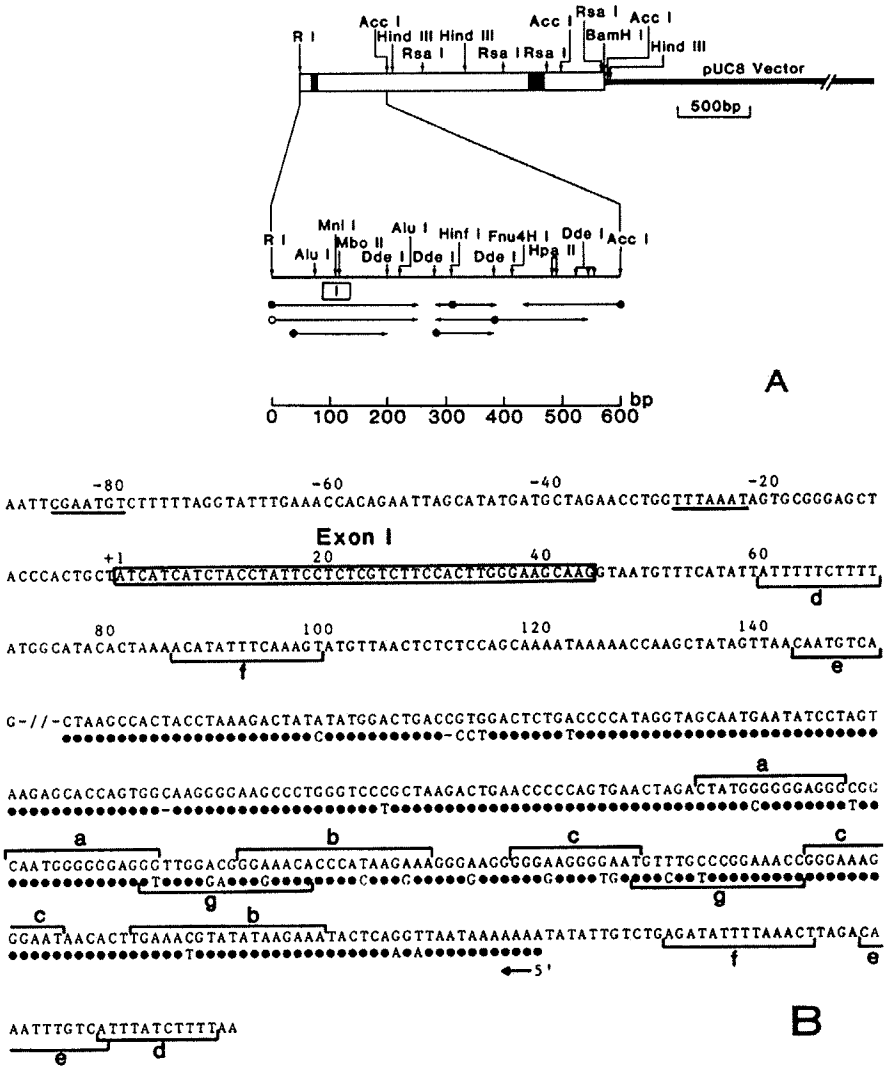

FIG. 8. **5′ sequence of the γ-casein gene.** *A*, Strategies used in sequencing the 5′ end of the γ-casein gene. Plasmid subclone pλγ2.1 DNA was digested either first with *Eco*RI, end-labeled, then with *Acc*I, or vice versa, purified from a 1% agarose gel, electroeluted, and sequenced from both ends of the 600-nucleotide fragment, expanded as shown. To sequence from the internal *Hinf*I or *Dde*I sites, the 600-nucleotide *Eco*RI/*Acc*I fragment was digested either with *Hinf*I, end-labeled, then with *Dde*I, or vice versa. *Arrows* with *solid circles* denote fragments labeled at the 3′ end with DNA polymerase I Klenow fragment; *arrows* with *open circles* denote fragments labeled at the 5′ end with T4 polynucleotide kinase. More than 80% of the sequence was determined from both directions. The position of the 5′ most exon, exon I, is indicated by an *open box*. *B*, 5′ nucleotide sequence of the rat γ-casein gene. The nucleotide sequences of the 0.6-kb *Eco*RI/*Acc*I fragment shown in Fig. *A*, which contains the 5′ flanking, first exon, and part of the first intervening sequence in the γ-casein gene, are shown in the 5′ to 3′ orientation. Exon I sequences are enclosed in the *rectangle*, with the position of the putative mRNA CAP site, A, designated as *+1*. The TATA sequence and a putative CAAT sequence upstream from the mRNA start site are *underlined*. The *double slash lines* starting at nucleotide 151 enclose an area of about 30 nucleotides whose sequence has not been determined. Solid circles indicate homology with a rat repeat clone (psubK) (Sargent, 1981). The psubK clone is about 4.3 kb long and contains a 2-kb repeat element (Sargent, 1981). *Dashes* indicate deletions. Sets of short repeat sequences are lettered with *a–g* (see Table I). The *arrow* indicates the orientation of the psubK clone.

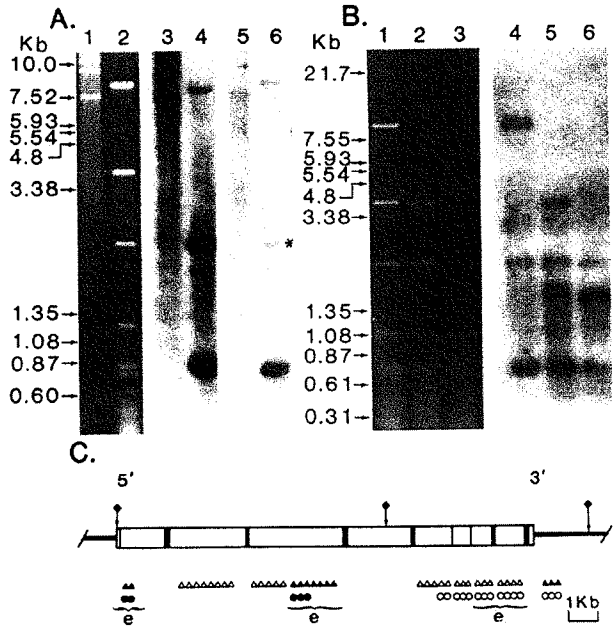

FIG. 9. **Localization of repeated sequences in the γ-casein gene.** *A*, λγ1 clone DNA was digested with *Eco*RI (*lane 1*) and *Msp*I plus *Eco*RI (*lane 2*) and run in duplicates (*lanes 3* and *4 versus lanes 5* and *6*). *B*, in another experiment, the same DNA was digested with *Msp*I plus the following: *Eco*RI (*lane 1*), *Bam*HI (*lane 2*), and *Sst*I (*lane 3*). The ethidium bromide patterns are shown on the *left* in *A* and *B*. After electrophoresing on a 1% agarose gel, the DNAs were hybridized with a ^{32}P-labeled total rat liver DNA probe at either 52 °C (*lanes 3* and *4* in *A*; *lanes 4–6* in *B*) or 68 °C (*lanes 5* and *6* in *A*). The temperature of the washes was the same as that used during hybridization. Molecular weight sizes are indicated by the *arrows*. *C*, a schematic diagram of a simplified γ-casein gene showing the locations of multiple repeated DNA sequences determined as in panels A and B, and numerous other experiments. Only the *Eco*RI sites are shown. △ and ▲, total rat liver DNA as probe; ○ and ●, psubK repeat DNA (Sargent, 1981) as probe (see Fig. 8*B*). The length of the symbols represents the shortest region of DNA known to contain the repeated sequences. *Solid versus open symbols* denote strong and weak hybridization signals. *e*, evolutionarily conserved repeats (Miesfeld *et al.*, 1981) (see text).

in pλγ2.2 (refer to Fig. 5 for locations) (data not shown). These results demonstrated that the large 15-kb γ-casein gene contains not only highly repeated sequences in the intron and flanking sequences, but also other sequences which are only found specifically within the γ-casein gene region.

DISCUSSION

The Rat γ-Casein Gene—The rat γ-casein gene exists in a single copy in the rat genome (Johnson *et al.*, 1983). Direct linkage of the γ-casein gene to the other members of this small gene family has yet to be demonstrated even after examining over 100 kb of rat DNA.[3] We have determined the structure of the γ-casein gene by examining a series of overlapping phage clones covering about 35 kb of the rat genome. One of the clones, λγ1, was found to contain the entire γ-casein gene. Using subclones of the λγ1 DNA, the ends of the γ-casein gene were defined by Southern hybridizations employing specific 5′ and 3′ cDNA probes as well as by direct sequencing of the 5′ end. The *Eco*RI digestion pattern of the γ-casein gene in total rat DNA revealed two fragments, which are slightly larger than those in the cloned λγ1 DNA (Fig. 3). The discrepancy is apparently not due to DNA rearrangement or polymorphisms, but may result instead from differential methylation around these *Eco*RI restriction enzyme sites in total rat DNA (see Johnson *et al.*, 1983).

The γ-casein gene is approximately 15 kb in length and contains a minimum of nine small exons separated by large intervening sequences as determined by R-loop analysis. It is possible that additional small exons may exist that are not detectable by either hybridization or R-loop analyses. For example, it has been observed recently by direct nucleotide sequencing that the β-casein gene contains exons as small as 24 nucleotides.[4] The presence of extremely small exons widely dispersed in the rat genome complicated the initial screening of the gene libraries even using full length cDNA probes. To alleviate this problem, large contiguous subgenomic fragments several kilobases in length and free of highly repeated sequences were used as probes in subsequent library screenings.

The 15-kb γ-casein gene is an unusually large and complex gene containing only 540 nucleotides of amino acid-encoding sequences and a total mRNA length of 869 nucleotides not including the poly(A) tail. Although there are other similarly large eucaryotic genes, e.g. the 15-kb rat serum albumin gene (Sargent *et al.*, 1981), 21-kb *Xenopus laevis* vitellogenin A1 gene (Wahli *et al.*, 1981), 37-kb chicken α2(I)collagen gene (Ohkubo *et al.*, 1980), and 25-kb hamster gene coding for carbamyl phosphate synthetase, aspartase transcarbamylase, and dihydroorotase (Padgett *et al.*, 1982), they also encode relatively large mRNAs, *i.e.* 2, 6, 4.6, and 7.9 kb, respectively, in size. Thus, these genes have relatively small intron/exon ratios ranging from 2 to 7:1 (Naora and Deacon, 1982). In contrast, the γ-casein gene has an intron/exon ratio of 16:1 which is greater than that predicted by Naora and Deacon (1982) after examining over 80 protein-encoding genes. However, the structure of the rat γ-casein gene is not unique, since the 31-kb mouse dihydrofolate reductase gene (Crouse *et al.*, 1982) has been reported to have an intron/exon ratio of 19:1.

DNA sequencing of the 5′ end of the γ-casein gene revealed the presence of a TATA sequence at nucleotide −29 which has also been found in a majority of other eucaryotic genes (Breathnach and Chambon, 1981). The sequence underlined at nucleotide −86 (Fig. 8*B*) is only weakly homologous to the concensus CAAT sequence (Benoist *et al.*, 1980), and the γ-casein gene may, therefore, not contain this sequence as has also been found in the mouse metallothionein (Glanville *et al.*, 1982), rat growth hormone (Barta *et al.*, 1981; Page *et al.*, 1981), and rat prolactin (Cooke and Baxter, 1982) genes. The first exon of the γ-casein gene consists of 44 nucleotides which encode 78% of the 5′ noncoding region of the γ-casein mRNA (Hobbs and Rosen, 1982). Interestingly, the 5′ noncoding sequences are highly conserved among the three rat casein mRNAs and may, therefore, represent a functional domain important for ribosomal binding and initiation of translation (Hobbs and Rosen, 1982). This split feature of the 5′ noncoding region has also been observed in a variety of other genes, *e.g.* the rat and chicken preproinsulin (Lomedico *et al.*, 1979; Perler *et al.*, 1980), rat calcitonin,[5] human α1-anti-trypsin (Leicht *et al.*, 1982), human pro-opiomelanocortin (Cochet *et al.*, 1982), chicken ovalbumin (Dugaiczyk *et al.*, 1979), and *Drosophila* actin (Fryberg *et al.*, 1981) genes.

The sequence of exon I agrees completely with the 5′ primer-extended sequence determined from reverse transcription of the γ-casein mRNA (Hobbs and Rosen, 1982). Assignment of the CAP site is tentatively placed on residue A based on comparisons with the nucleotide sequences around the mRNA start sites in both α- and β-casein mRNAs (Hobbs and Rosen, 1982), which begin with 5′-AUC. Furthermore, residue A has been shown to be the preferred CAP site for over 60 mRNAs (Breathnach and Chambon, 1981). Moreover, no canonical splice junction sequences (Breathnach and Chambon, 1981) are observed immediately upstream in the 5′ flanking region. Definitive assignment of the transcriptional start site can be obtained only by S1 nuclease mapping (Berk and Sharp, 1977).

An interesting although hypothetical intrastrand stem-loop structure can be formed at the 5′ end of the gene. A 19-bp stem structure can be formed with sequences from nucleotides −31 to −13, which include the TATA signal, and nucleotides 111–129 in the intron (74% homology), such that the first exon is enclosed within a 122-nucleotide hairpin loop (refer to Fig. 8*B*; data not shown). However, until the 5′ genomic sequences of the α- and β-casein genes are available for comparison, it is premature to speculate the significance of this structure, or whether exon I may have an independent origin much like the first exon found in the rat prolactin (Cooke and Baxter, 1982), rat growth hormone (Barta *et al.*, 1981; Page *et al.*, 1981), and chicken conalbumin (Cochet *et al.*, 1979) genes. It is also possible that this intrastrand loop may form a single-stranded S1 nuclease-sensitive region which may be involved in the initiation of transcription of active genes as recently proposed by Larsen and Weintraub (1982) and McKnight (1982). Since only the first exon and part of the first intron have been sequenced, it has not yet been determined whether the remainder of the exons in the γ-casein gene are positioned corresponding to potential functional domains as observed in several other genes (Artymiuk *et al.*, 1981; Craik *et al.*, 1980; Eaton, 1980; Stein *et al.*, 1980).

Repeated Sequences in and around the γ-Casein Gene—Several families of repeated DNA sequences were observed in the γ-casein gene, both in the intervening and flanking regions. These repeats were detected by nick translating total rat liver DNA as probe; and, as estimated by Shen and Maniatis (1980), only those repeated DNA sequences in the probe which are present in greater than 50 copies/rat genome will be able to generate a hybridization signal. These highly repeated sequences in the γ-casein gene most likely belong to the short interspersed repeated DNA families found commonly in between and within genes in mammalian genomes (Jagadeeswaran *et al.*, 1981; Singer, 1982). The different fam-

[3] L.-Y. Yu-Lee, unpublished observations.
[4] W. K. Jones, unpublished observations.
[5] G. Rosenfeld, personal communication.

ilies of repeats may be due to differences in repetition frequency, sequence length, or sequence homology. In addition to the highly repeated DNA, some of the repeated sequences appear to be reiterated only in the γ-casein gene region, such as those found in the 2.1- and 4.1-kb subgenomic fragments at the 5′ end of the gene (data not shown). Similar findings of repeats belonging to the same family but dispersed throughout a gene region have been reported for mouse serum albumin and α-fetoprotein (Kioussis *et al.*, 1981), vitellogenin (Ryffel *et al.*, 1981), fibroin (Pearson *et al.*, 1981), and rabbit β-globin (Shen and Maniatis, 1980) genes. Adding to the complexity of the repetitive elements in the γ-casein gene, three intron regions are found to contain evolutionarily conserved repeats which are conserved among human, slime mold, and yeast DNAs (Miesfield *et al.*, 1981). These repeats are distinct from the ubiquitous *Alu* family of repeated sequences found interspersed with single copy DNA in human (Houck *et al.*, 1979; Jelinek *et al.*, 1980; Schmid and Jelinek, 1982) and other eucaryotic DNAs including mouse (Krayev *et al.*, 1980) and Chinese hamster (Haynes *et al.*, 1981) DNA. Finally, at least three families of repeated sequences associated with the rat ribosomal genes (Mroczka *et al.*, 1982) have also been localized in and around the γ-casein gene.[6]

A 282-nucleotide long repeated DNA sequence found in the first intron is observed to share a 92% homology with a repetitive element found in the 3′ flanking region of the rat serum albumin gene (Sargent, 1981). This albumin gene-associated rat repeat has also been localized at the 3′ end of the rat seminal vesicle secretory protein IV gene[7] as well as near several rat ribosomal genes.[6] However, since this repeat sequence is present in 100,000 copies in the rat genome (Sargent, 1981), it is not surprising to have found this repeat interspersed throughout the intron and flanking regions of the large 15-kb γ-casein gene.

The γ-casein intron I repeat is rich in purine nucleotides and is in part composed of a series of internally duplicated repeated sequences as has been also found for the human *Alu* repetitive sequences (Schmid and Jelinek, 1982) and the rat 93-bp highly repetitive DNA families (Sealy *et al.*, 1981). Table I (Internal repeats) shows three sets of short internally repeated sequences (with 72–93% homology) found within the γ-casein intron I repeat element. These repeat sequences correspond to the unique, albeit, 3′ flanking sequences located near the rat albumin gene. Several direct repeats of 8–14 bp (with 71–91% homology) have been found to flank the intron I repeat as shown in Table I (Flanking repeats). It is not known whether this repetitive element found associated with both the γ-casein and albumin genes is the result of a gene conversion event (Egel, 1981) between the two rat genes which are located on the same chromosome (Gupta *et al.*, 1982) or more likely a transposition (Calos and Miller, 1980) of the repeat sequence into the γ-casein gene. These short direct repeats could conceivably be vestiges of an insertion event similar to the insertion of an *Alu*-like repeat sequence into the rat growth hormone gene (Barta *et al.*, 1981; Payvar *et al.*, 1981), rat α-tubulin pseudogene (Lemischka and Sharp, 1982), and mouse pro-α1(I)procollagen gene (Monson *et al.*, 1982).

Another interesting feature in the intron I repeat region is a potential hairpin loop of 29 nucleotides, generated by a 13-bp inverted repeat with sequences denoted by the *g* in Fig. 8*B*. Whether or not such a loop structure occurs *in vivo* is not known. However, interestingly, this putative loop structure contains two *Hpa*II/*Msp*I restriction recognition sequences, 5′-CCGG-3′, in its stem. One of these *Hpa*II sites is variably methylated, and its increased methylation correlates with the lowered expression of the γ-casein gene in tumor tissues (see Johnson *et al.*, 1983).

TABLE I

Short repeats within and flanking the 282-bp repetitive element in intron I

Solid circles indicate homology between the short repeats which are denoted by sets of letters a–f in Fig. 8*B*.

		Homology
		%
Internal repeats		
	CTATGGGGGGAGGG	
(a)	•A••••••••••••	93
	GGAAACACCCATAAGAAA	
(b)	T•••••GTAT•••••••••	72
	GGAAGGGGAAT	
(c)	••••A•••••••	92
Flanking repeats		
	ATTTTTCTTTT	
(d)	••••A•••••••	91
	CAATGTCA	
(e)	••••••••	73
	ATT	
	ACATATTTCAAAGT	
(f)	•G••••••T•••C•	78

Although the functions of repetitive DNA are unknown, it has been proposed by Gilbert (1978) that dispersed repetitive sequences may serve as "hot spots" for genetic recombination and that their presence in introns may be responsible for exon shuffling and gene rearrangement (Darnell, 1978). In this regard, it is interesting to note that the albumin gene-associated repetitive element (Sargent, 1981) found in the first intron of the γ-casein gene has also been localized to the 5′ end, as well as other regions, of the β-casein gene.[8] Moreover, it has been shown that the γ- and β-casein genes do not cross-hybridize through their structural gene sequences (Richards *et al.*, 1981b) but rather only through these repeated sequences in the various introns.[8]

Having characterized the genomic γ-casein gene sequences, it may now be possible to define the transcription unit of the γ-casein gene. Unique genomic DNA fragments can be employed to provide long contiguous repeat-free DNA probes for studies of RNA processing. These subgenomic probes can also be used to examine the mechanisms of hormone action by localizing regulatory sequences via gene transfection experiments (Gluzman, 1982) as well as by direct hormone binding studies (Payvar *et al.*, 1981; Pfahl, 1982). Finally, in the accompanying paper (Johnson *et al.*, 1983), characterization of the γ (this paper)- and β-casein gene[8] structures allowed the detailed analysis of the inverse relationship between certain sites of DNA methylation and casein gene expression.

Acknowledgments—We wish to thank Dr. Donald A. Richards for his assistance in library screening, Dr. Miles Mace, Jr. for his assistance in R-loop analysis, Drs. Tom Sargent, Linda Jagodzinski, and James Bonner for providing the rat DNA libraries, and Patricia Kettlewell for preparation of this manuscript.

REFERENCES

Artymiuk, P. J., Blake, C. C. F., and Sippel, A. E. (1981) *Nature (Lond.)* **290**, 287–288

[6] L.-Y. Yu-Lee, D. Mroczka, L. Rothblum, and J. M. Rosen, unpublished observations.

[7] S. E. Harris, personal communciation.

[8] L.-Y. Yu-Lee, S. M. Clift, W. K. Jones, and J. M. Rosen, manuscript in preparation.

Barta, A., Richards, R. I., Baxter, J. D., and Shine, J. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 4867–4871
Benoist, C., O'Hare, K., Breathnach, R., and Chambon, P. (1980) *Nucleic Acids Res.* **8,** 127–142
Benton, W. D., and Davis, R. W. (1977) *Science (Wash. D. C.)* **196,** 180–182
Berk, A. J., and Sharp, P. A. (1977) *Cell* **12,** 721–732
Blackburn, D. E., Hobbs, A. A., and Rosen, J. M. (1982) *Nucleic Acids Res.* **10,** 2295–2307
Blattner, F. R., Williams, B. G., Blechl, A. E., Denniston-Thompson, K., Farber, H. E., Furlong, L.-A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L., and Smithies, O. (1977) *Science (Wash. D. C.)* **196,** 161–169
Bolivar, F. (1978) *Gene (Amst.)* **4,** 121–136
Breathnach, R., and Chambon, P. (1981) *Annu. Rev. Biochem.* **50,** 349–383
Calos, M. P., and Miller, J. H. (1980) *Cell* **20,** 579–595
Chaconas, G., and van de Sande, J. H. (1980) *Methods Enzymol.* **65,** 75–85
Cochet, M., Gannon, G., Hen, R., Maroteaux, L., Perrin, F., and Chambon, P. (1979) *Nature (Lond.)* **282,** 567–574
Cochet, M., Chang, A. C. Y., and Cohen, S. N. (1982) *Nature (Lond.)* **297,** 335–339
Cooke, N. E., and Baxter, J. D. (1982) *Nature (Lond.)* **297,** 603–606
Craik, C. S., Buchman, S. R., and Beychok, S. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77,** 1384–1388
Crouse, G. F., Simonsen, C. C., McEwan, R. N., and Schimke, R. T. (1982) *J. Biol. Chem.* **257,** 7887–7897
Dagert, M., and Ehrlich, S. D. (1979) *Gene (Amst.)* **6,** 23–28
Darnell, J. E., Jr. (1978) *Science (Wash. D. C.)* **202,** 1257–1259
Denhardt, D. T. (1966) *Biochem. Biophys. Res. Commun.* **23,** 641–646
Dretzen, G., Bellard, M., Sassone-Corsi, P., and Chambon, P. (1981) *Anal. Biochem.* **112,** 295–298
Dugaiczyk, A., Boyer, H. W., and Goodman, H. M. (1975) *J. Mol. Biol.* **96,** 171–184
Dugaiczyk, A., Woo, S. L. C., Colbert, D. A., Lai, E. C., Mace, M. L., Jr., and O'Malley, B. W. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76,** 2253–2257
Eaton, W. A. (1980) *Nature (Lond.)* **284,** 183–185
Egel, R. (1981) *Nature (Lond.)* **290,** 191–192
Esumi, H., Takahashi, Y., Sato, S., and Sugimura, T. (1982) *Nucleic Acids Res.* **10,** 4247–4257
Fitzgerald, M., and Shenk, T. (1981) *Cell* **24,** 251–260
Fryberg, E. A., Bond, B. J., Hershey, N. D., Mixter, K. S., and Davidson, N. (1981) *Cell* **24,** 107–116
Gilbert, W. (1978) *Nature (Lond.)* **271,** 501
Glanville, N., Durnam, D. M., and Palmiter, R. D. (1981) *Nature (Lond.)* **292,** 267–269
Gluzman, T. (1982) *Eukaryotic Viral Vectors,* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
Griffin-Shea, R., Thireos, G., Kafatos, F. C., Petri, W. H., and Villa-Komaroff, L. (1980) *Cell* **19,** 915–922
Grosclaude, F., Mercier, J. C., and Dumas, B. R. (1973) *Neth. Milk Dairy J.* **27,** 328–340
Gupta, P., Rosen, J. M., D'Eustachio, P., and Ruddle, F. H. (1982) *J. Cell Biol.* **93,** 199–204
Guyette, W. A., Matusik, R. J., and Rosen, J. M. (1979) *Cell* **17,** 1013–1023
Haynes, S. R., Toomey, T. P., Leinwald, L., and Jelinek, W. R. (1981) *Mol. Cell. Biol.* **1,** 573–583
Hobbs, A. A., Richards, D. A., Kessler, D. J., and Rosen, J. M. (1982) *J. Biol. Chem.* **257,** 3598–3605
Hobbs, A. A., and Rosen, J. M. (1982) *Nucleic Acids Res.* **10,** 8079–8098
Holmes, D., and Quigley, M. (1981) *Anal. Biochem.* **114,** 193–197
Houck, C. M., Rinehart, F. P., and Schmid, C. W. (1979) *J. Mol. Biol.* **132,** 289–306
Jagadeeswaran, P., Forget, B. G., and Weissman, S. M. (1981) *Cell* **26,** 141–142
Jelinek, W. R., Toomey, T. P., Leinwald, L., Duncan, C. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L., and Schmid, C. W. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77,** 1398–1402
Johnson, M. L., Levy, J., Supowit, S. C., Yu-Lee, L.-Y., and Rosen, J. M. (1983) *J. Biol. Chem.* **258,** 10805–10811
Katz, L., Kingsbury, D. T., and Helinski, D. R. (1973) *J. Bacteriol.* **114,** 577–591
Kioussis, D., Eiferman, F., van de Rijn, P., Gorin, M. B., Ingram, R. S., and Tilghman, S. M. (1981) *J. Biol. Chem.* **256,** 1960–1967
Krayev, A. S., Kramerov, D. A., Skryabin, K. G., Ryskov, A. P., Bayev, A. A., and Georgiev, G. P. (1980) *Nucleic Acids Res.* **8,** 1201–1215
Larsen, A., and Weintraub, H. (1982) *Cell* **29,** 609–622
Leder, P., Tiemeier, D., and Enquist, L. (1977) *Science (Wash. D. C.)* **196,** 175–177
Leicht, M., Long, G. L., Chandra, T., Kurachi, K., Kidd, V. J., Mace, M., Jr., Davie, E. W., and Woo, S. L. C. (1982) *Nature (Lond.)* **297,** 655–659
Lemischka, I., and Sharp, P. A. (1982) *Nature (Lond.)* **300,** 330–335
Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R., and Tizard, R. (1979) *Cell* **18,** 545–558
Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K., and Efstratiadis, A. (1978) *Cell* **15,** 687–701
Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual,* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
Maxam, A. M., and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 560–564
Maxam, A. M., and Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560
McKnight, S. L. (1982) *Cell* **31,** 355–365
Miesfeld, R., Krystal, M., and Arnheim, N. (1981) *Nucleic Acids Res.* **9,** 5931–5947
Monson, J. M., Friedman, J., and McCarthy, B. J. (1982) *Mol. Cell. Biol.* **2,** 1362–1371
Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472
Mroczka, D. L., Cassidy, B. C., Yu-Lee, L. Y., Hawkins, J. W., and Rothblum, L. I. (1982) *J. Cell. Biol.* **95,** 212a (abst.)
Naora, H., and Deacon, N. J. (1982) *Proc. Natl. Acad. Sci. U. S. A.* **79,** 6196–6200
Norgard, M. V., Keem, K., and Monahan, J. J. (1978) *Gene (Amst.)* **3,** 279–292
Norgard, M. V., Emigholz, K., and Monahan, J. J. (1979) *J. Bacteriol.* **138,** 270–272
Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, V. B., Sullivan, M., Pastan, I., and de Crombrugghe, B. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77,** 7059–7063
Padgett, R. A., Wahl, G. M., and Stark, G. R. (1982) *Mol. Cell. Biol.* **2,** 302–307
Page, G. S., Smith, S., and Goodman, H. M. (1981) *Nucleic Acids Res.* **9,** 2087–2104
Payvar, F., Wrange, O., Carlstedt-Duke, J., Okret, S., Gustafsson, J. A., and Yamamoto, K. R. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 6628–6632
Pearson, W. R., Mukai, T., and Morrow, J. F. (1981) *J. Biol. Chem.* **256,** 4033–4041
Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J. (1980) *Cell* **20,** 555–566
Pfahl, M. (1982) *Cell* **31,** 475–482
Prentki, P., Karch, F., Iida, S., and Meyer, J. (1981) *Gene (Amst.)* **14,** 289–299
Proudfoot, N. J., and Brownlee, G. G. (1976) *Nature (Lond.)* **263,** 211–214
Richards, D. A., Rodgers, J. R., Supowit, S. C., and Rosen, J. M. (1981a) *J. Biol. Chem.* **256,** 526–532
Richards, D. A., Blackburn, D. E., and Rosen, J. M. (1981b) *J. Biol. Chem.* **256,** 533–538
Rigby, P. W. J., Dickmann, M., Rhodes, C., and Berg, P. (1977) *J. Mol. Biol.* **112,** 237–251
Rosen, J. M. (1976) *Biochemistry* **15,** 5263–5271
Rosen, J. M., Woo, S. L. C., and Comstock, J. P. (1975) *Biochemistry* **4,** 2895–2903
Rosen, J. M., Richards, D. A., Guyette, W., and Matusik, R. J. (1980) in *Gene Regulation by Steroid Hormones* (Roy, A., and Clark, J., eds) pp. 58–77, Springer-Verlag, New York
Rosen, J. M., and Shields, D. (1980) in *Testicular Development, Structure and Function* (Steinberger, A., and Steinberger, E., eds) pp. 343–349, Raven Press, New York
Ryffel, G. U., Muellener, D. B., Wyler, T., Wahli, W., and Weber, R. (1981) *Nature (Lond.)* **291,** 429–431
Sargent, T. D. (1981) Ph.D. dissertation, Stanford University
Sargent, T. D., Wu, J.-R., Sala-Trepat, J. M., Wallace, R. B., Reyes, A. A., and Bonner, J. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76,** 3256–3260
Sargent, T. D., Jagodzinski, L. L., Yang, M., and Bonner, J. (1981) *Mol. Cell. Biol.* **1,** 871–883
Schmid, C. W., and Jelinek, W. R. (1982) *Science (Wash. D. C.)* **216,**

1065–1070
Sealy, L., Hartley, J., Donelson, J., Chalkey, R., Hutchinson, N., and Hamkalo, B. (1981) *J. Mol. Biol.* **145,** 291–318
Shen, J. C. K., and Maniatis, T. (1980) *Cell* **19,** 379–391
Singer, M. F. (1982) *Cell* **28,** 433–434
Smith, A. J. H. (1980) *Methods Enzymol.* **65,** 560–580
Smith, G. E., and Summers, M. D. (1980) *Anal. Biochem.* **109,** 123–129
Smith, H. O., and Birnsteil, M. L. (1976) *Nucleic Acids Res.* **3,** 2387–2398
Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517
Stein, J. P., Catterall, J. F., Kristo, P., Means, A. R., and O'Malley, B. W. (1980) *Cell* **21,** 681–687
Topper, Y. J. (1970) *Recent Prog. Horm. Res.* **26,** 286–308
Vieira, J., and Messing, J. (1982) *Gene (Amst.)* **19,** 259–268
Wahl, G. M., Stern, M., and Stark, G. R. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76,** 3683–3687
Wahli, W., Dawid, I. B., Ryffel, G. U., and Weber, R. (1981) *Science (Wash. D. C.)* **212,** 298–304
Waugh, D. F. (1971) in *Milk Proteins* (McKenzie, H. A., ed) Vol. II, pp. 3–85, Academic Press, New York
Woo, S. L. C., Dugaiczyk, A., Tsai, M.-J., Lai, E. C., Catteral, J. F., and O'Malley, B. W. (1978) *Proc. Natl. Acad. Sci. U. S. A.* **75,** 3688–3692